PAPER

Reliability of the variables in a new set of models that predict outcome after stroke

N U Weir, C E Counsell, M McDowall, A Gunkel, M S Dennis

J Neurol Neurosurg Psychiatry 2003;**74**:447–451

See end of article for authors' affiliations

Correspondence to: Dr M Dennis, Department of Clinical Neurosciences, Western General Hospital, Edinburgh EH4 2XU, UK; msd@skull.dcn.ed.ac.uk

Received 2 October 2002
Accepted 13 November 2002

Objectives: To provide valid predictions of outcome, the variables included in a prognostic model must be capable of reliable collection. The authors have recently reported a set of simple but rigorously developed models that predict outcome after stroke. The aim of this study was to establish the inter-rater reliability of the variables included in the models.

Methods: Inter-rater agreement was measured prospectively (between two clinicians; 92 patients) and retrospectively (between two auditors; 200 patients) and the validity of the data collected retrospectively was estimated by comparing them with data collected prospectively (195 patients). In the prospective study inter-rater agreement for urinary incontinence and for the variables of three other previously published models was also measured. The median difference (md) between ages and κ statistics for other variables was calculated.

Results: For the model variables, prospective agreement ranged from good to excellent (age: md 0 years; living alone before the stroke κ 0.84; pre-stroke functional independence κ 0.67; normal verbal Glasgow Coma Scale score κ 0.79; ability to lift both arms against gravity κ 0.97; ability to walk unaided κ 0.91) while retrospective agreement (age: md 0 years; κ 0.55–0.92) and agreement between prospective and retrospective observers (age: md 0 years; κ 0.49–0.78) was acceptable but less good. Prospective agreement was excellent for urinary incontinence (κ 0.87) and variable for the other models (κ 0.23–0.81)

Conclusion: The variables included in these new simple models of outcome after stroke are capable of reliable collection, comparable to or better than that of the other predictive variables considered. When collected retrospectively, the model variables are likely to remain reliable and reasonably valid.

Accurate predictions of outcome made soon after the onset of stroke have a number of important applications, such as: informing communication with patients and relatives; supporting treatment decisions; improving stratification of patients in randomised controlled trials; and improving comparisons of observational data by allowing for better adjustment for casemix. Unfortunately, despite many attempts to develop statistical models predicting outcome after stroke, none have achieved widespread acceptance, partly because none have been rigorously developed.[1 2] One important aspect of quality that has often been overlooked by those developing models is the reliability of the predictive variables—that is, how reproducible the variables are when measured again by the same person (intra-rater reliability) and when measured by two or more different people (inter-rater reliability).[2–4] As prognostic models may be used by a wide variety of different people, inter-rater reliability is particularly important.[3] The few data that do describe the inter-rater reliability of predictive variables for patients with stroke mostly relate to items included in the standard neurological examination or in certain stroke scales.[5–7] When the reliability of variables included in prognostic models has been studied, some have been found to include variables with poor inter-rater reliability, for example, the Mathew score, included in the Uppsala model, has a low inter-rater reliability in patients with stroke.[8 9] Furthermore, many existing models were developed from or are applied using data collected from the medical record.[2] Retrospective data such as these may be less reliable and less accurate than prospectively collected data and might be expected to result in flawed models or inaccurate predictions of outcome.[10–12] However, the inter-rater reliability and the accuracy of retrospectively collected predictive data for patients with stroke has been little studied.[13–17]

We have recently reported a set of prognostic models for patients with acute and sub-acute stroke.[18] Each model is based on the same six simple clinical variables (age; living alone before the stroke; pre-stroke functional independence; normal verbal Glasgow Coma Scale score; ability to lift both arms against gravity; ability to walk unaided), all of which can be collected at the patient's bed side. We developed the models according to established guidelines using a training dataset taken from the Oxfordshire Community Stroke Project and have shown that they predict survival and functional status accurately in two large and independent cohorts.[18] The models have already been used to adjust for differences in casemix between cohorts of patients managed in different hospitals and to stratify patients in clinical trials on the basis of baseline predicted risk.[19–22] As such, our models are not only practical and widely applicable but also the most rigorously developed and tested to date. The aim of this study was to determine the inter-rater reliability of the variables included in our models. We estimated their inter-rater reliability when collected prospectively (at the patient's bed side) and when collected retrospectively (from the medical record). We compared the inter-rater reliability of our prospectively collected model variables with that of urinary incontinence (probably the single most important predictive variable in stroke)[23 24] and the variables of three previously published models.[18 25–27] We also estimated the accuracy of our retrospectively collected data by comparing them with prospectively collected data.

METHODS

We tested the reliability of data collection in three different ways and used a different group of patients for each:

Table 1 Definitions of the variables collected

Variable	Definition	Unassessable data
Our simple models [18]		
Age	(Date of stroke onset - date of birth) (y)	Not applicable
Living alone pre-stroke	No other person living permanently with the person before the stroke	Coded yes
Independent pre-stroke	Independent in activities of daily living (Oxford Handicap Score < 3) before stroke	Coded yes
Normal verbal GCS	Patient oriented in time, place and person	Coded no
Lift both arms	Able to lift both arms against gravity to 90 degrees (that is, MRC score 3 or more)	Coded no
Able to walk	Able to walk without the help of another person (can use walking aid if needed)	Coded no
Other models and variables		
Urinary incontinence	Incontinent of urine since stroke or catheterised or penile sheath	Coded no
Guy's model [25]		
Loss of consciousness at onset	Loss of consciousness at onset of stroke	Coded no
Drowsy	Drowsy or comatose	Not applicable
Complete limb paralysis	Power in worst limb MRC 0 or 1	Coded no
TACS	Hemiplegia and hemianopia and either dysphasia or other higher cortical dysfunction	Not applicable
Motor LACS	Motor hemideficit with or without sensory loss but no hemianopia or higher cortical features	Not applicable
Edinburgh model [26]		
Arm function	0=No weakness 1=Able to lift to shoulder height and resist active movement but weak 2=Able to lift to shoulder height but cannot resist active movement 3=Able to move arm but cannot lift it to shoulder height 4=No flicker of contraction in affected arm	Coded as 4
Arm proprioception	0=Finds thumb of affected arm with other hand with eyes closed accurately first time 1=Finds thumb within 5 seconds having missed it initially 2=Finds affected arm and then follows this up to affected thumb 3=Unable to find thumb	Coded as 3
Postural capability	0=Able to walk without human assistance for about 10 feet 1=Able to maintain standing position without support but cannot walk without human assistance 2=Able to sit with legs over side of bed without support but cannot stand 3=Unable to sit without help	Coded as 3
Orpington model [27]*		
Abbreviated Mental Test Score	0=score 10 1=score 8 to 9 2=score 5 to 7 3=score 0 to 4	Coded as 3

MRC, Medical Research Council; BP, blood pressure; GCS, Glasgow coma score; TACS, total anterior circulation stroke; LACS, lacunar stroke. *The Orpington model comprises the variables of the Edinburgh model plus the Abbreviated Mental Test score.

Prospective clinical assessment

We measured agreement between prospective observers in a consecutive cohort of 92 patients with an acute stroke admitted to the Western General Hospital, Edinburgh between March 1997 and September 1997. Two neurology trainees (NW and CC) examined each patient on the same day, blind to the findings of the other and collected data on several predictive variables (table 1). Where possible, the definitions of the variables of the previously published models were taken from the original papers. Information on each variable was collected from the patient themselves or, where necessary (for example, if the patient was confused or dysphasic), by interviewing the relatives and searching the hospital notes.

Retrospective data collection from routine medical records

We measured agreement between retrospective observers in 200 patients with an acute stroke admitted to any of five Scottish hospitals between August 1995 and July 1997 and who were included in a previously reported study.[19] A neurology trainee (NW) and an audit assistant (AG) independently extracted the predictive variables from the medical record pertaining to the day of admission (including the nursing entries).

Retrospective compared with prospective data collection

We measured the agreement between retrospective and prospective data collection in 195 patients with an acute stroke admitted to the Western General Hospital, Edinburgh between August 1995 and July 1997 and who were included in both the previously reported study[19] and our prospective stroke register. One of five neurology trainees collected the predictive variables at the patient's bed side on the day or day after admission. We compared these data with those abstracted by an audit assistant (AG) from the patient's medical record (including the nursing entries) pertaining to the day of admission.

Statistical analyses

We measured agreement using the methods suggested by Altman.[28] For age, we calculated the median difference between the observers with 5th to 95th centiles. For categorical variables we calculated the simple proportion of agreement between observers, the κ value and the 95% confidence intervals[29] and, where appropriate, a weighted κ. The κ value describes agreement beyond chance and, in general, κ values of 0 to 0.20 indicate poor agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 good agreement, and 0.81 to 1.00 excellent agreement.[28] Systematic disagreement can be identified by inspecting the data in a contingency table, its presence being revealed by imbalance in the "off diagonal" cells. We estimated the significance of any such suspected bias using McNemar's test.[30] We performed all calculations using SPSS (version 6.1 for Windows) and the Confidence Interval Analysis program (version 1.0).

RESULTS

Prospective clinical assessment

The median interval between stroke onset and clinical assessment was one day (interquartile range 0 to 3 days). The mean delay between the two assessments (by NW and CC) was 3.5 hours (SD 1.5 hours). We collected data on urinary

Table 2 Inter-rater agreement in the three different studies

Variable	Prospective study (n=92)			Retrospective study (n=200)		Prospective *v* retrospective study (n=195)	
	% agreement	κ (95% CI)	Weighted κ	% agreement	κ (95% CI)	% agreement	κ (95% CI)
Our simple models							
Lived alone pre-stroke	92	0.84 (0.72 to 0.95)	–	97	0.92 (0.87 to 0.98)	90	0.78 (0.68 to 0.87)
Independent pre-stroke (ADL)	88	0.67 (0.50 to 0.85)	–	90	0.61 (0.46 to 0.76)	86	0.49 (0.33 to 0.65)
Normal verbal GCS	90	0.79 (0.66 to 0.92)	–	86*	0.73 (0.63 to 0.82)	83 †	0.60 (0.48 to 0.73)
Lift both arms	99	0.97 (0.91 to 1.00)	–	94	0.88 (0.81 to 0.95)	89	0.71 (0.59 to 0.83)
Able to walk	96	0.91 (0.83 to 1.00)	–	87	0.55 (0.40 to 0.70)	81	0.61 (0.50 to 0.73)
Other models and variables							
Urinary incontinence	94	0.87 (0.76 to 0.98)	–	–	–	–	–
Guy's model							
Loss of consciousness	93	0.59 (0.30 to 0.88)	–	–	–	–	–
Drowsy	84	0.52 (0.31 to 0.73)	–	–	–	–	–
Complete limb paresis	93	0.81 (0.66 to 0.96)	–	–	–	–	–
TACS	89	0.23 (0.00 to 0.55)	–	–	–	–	–
Motor LACS	88	0.63 (0.42 to 0.83)	–	–	–	–	–
Edinburgh model							
Arm function	72	0.53 (0.39 to 0.67)	0.72	–	–	–	–
Arm proprioception	60	0.39 (0.26 to 0.53)	0.53	–	–	–	–
Postural capability	85	0.76 (0.65 to 0.87)	0.84	–	–	–	–
Orpington model ‡							
Abbreviated Mental Test Score	60	0.45 (0.31 to 0.59)	0.64	–	–	–	–

ADL, activities of daily living; GCS, Glasgow coma score; TACS, total anterior circulation stroke; LACS, lacunar stroke. *n=199. †n=194. ‡The Orpington model comprises the variables of the Edinburgh model plus the Abbreviated Mental Test score.

incontinence in only 86 patients. The median difference in the assessment of age between the two observers was zero years (5th to 95th centiles 0 to 0 years). The observers agreed precisely on age in 80 patients (87%) and disagreed in 10 patients by up to four days and in two patients by up to two years (the differences were attributable to discrepancies between the medical notes and patient or because confused patients gave different dates of birth). The inter-rater reliability of the other predictive factors are shown in table 2. We achieved good to excellent agreement for the five categorical variables included in our own simple models. Of these, the lowest level of agreement (κ 0.67) was for pre-stroke independence in the activities of daily living (disagreement was partly systematic (z 2.41, p 0.016); with NW less likely to judge patients independent than CC). We achieved excellent agreement on urinary incontinence and moderate to excellent agreement on all other predictive variables assessed, except for the identification of "total anterior circulation stroke" on which agreement was only fair. Taken as a set, we achieved a higher level of agreement for the predictive variables included in our own models than for those included in the three previously published models.

Retrospective data extraction from routine medical records

The median difference in the assessment of age between observers was zero years (5th to 95th centiles 0 to 0 years). The observers agreed precisely on age in 194 cases (97%) and differed by one year in four cases, two years in one case, and 10 years in one case (discrepancies were attributable to transcription error by the observers and to differences in dates in different parts of the medical record). The inter-rater reliability of the remaining variables in our simple models ranged from good to excellent except for the ability to walk unaided where agreement was "upper" moderate (κ 0.55) (table 2). For pre-stroke independence and the verbal GCS score disagreement was partly systematic (z 2.01, p 0.044 and z 2.89, p 0.004, respectively); in each case, NW was less likely to judge the patient to be independent or normal than AG.

Retrospective versus prospective collection

We collected data prospectively on the day of admission in 35 patients and on the day after admission in 160 patients. The median difference in the assessment of age between retrospective and prospective observers was zero years (5th to 95th centiles 0 to 0 years). The observers agreed precisely on age in 186 cases (95%) and differed by one to three days in eight cases and seven months in one case. The inter-rater reliability (the validity) of extracting the categorical variables of our simple models from the medical record ranged from moderate to excellent (table 2), although the level of agreement was always less than that achieved by two prospective observers. Disagreement on pre-stroke functional independence was partly systematic (z 4.23, $p<0.0001$), with the retrospective observer tending to judge the patient to be independent when the prospective observers did not.

DISCUSSION

This study shows that the six variables included in our simple models of outcome after stroke can be collected very reliably by clinicians at the bed side. It also suggests that, when collected prospectively, the inter-rater reliability of our

variables is comparable to that of urinary incontinence and comparable to or better than that of the variables included in the three previously published models. Furthermore, when collected retrospectively, the variables included in our models remained reliable and reasonably valid. These findings further enhance the validity of our models and suggest that they can be successfully applied not only in clinical and research settings (where prospective data collection is likely) but also in the field of audit and quality control (where retrospective data collection is more often the case).

The satisfactory reliability of the variables in our models is likely to reflect our decision to exclude, as far as was possible, those variables with known or presumed low reliability (for example, sensory impairments) and variables that are informative in only a small proportion of patients (for example, bilateral extensor plantar reflexes) during model development.[18] It is notable that the less reliable variables included in the other models that we studied were often complex or required skilled interpretation of clinical findings, or both. The Edinburgh and Orpington models also include variables with three or more categories and such variables always have lower κ values than dichotomous variables.[28] The lower inter-rater reliability of some of the variables in the other models may partly explain their poor performance when tested in independent cohorts.[23]

Of our six variables, we achieved the lowest level of inter-rater agreement over the three assessments for the variable describing functional independence in activities of daily living before the stroke (κ 0.49–0.67). In each assessment, disagreement between observers was partly systematic. Discussion revealed that this was because of minor variation between observers in the definition of activities of daily living. More reliable assessments of functional independence might be possible if a checklist were used to specify the ADLs that should be considered and the threshold at which the patient should be considered dependent. While ADLs are often taken to include washing, dressing, feeding, toileting, and mobilising,[31] it is less clear, for instance, whether bathing or shopping should be included as these are not necessarily daily activities. A definition of functional independence that excludes bathing and shopping would probably be sensible given the importance of environmental factors in determining abilities in these areas (bath or shower; distance from shops).

Comparisons of inter-rater reliability data between different populations should be performed cautiously as the level of agreement achieved between observers is partly governed by the prevalence of the attribute within each population.[28] None the less, as might have been expected, we found generally better agreement between observers when data were collected prospectively than when they were collected retrospectively. In particular, the ability to walk unaided was extremely reliable when assessed prospectively but only moderately reliable when assessed retrospectively. Reviewing the hospital notes showed that this discrepancy was probably because of the infrequency with which physicians specifically record the ability to walk soon after admission. These findings support the idea that if models such as ours are to be used routinely (for example, to adjust for differences in casemix [19]) it would be preferable for the clinicians to explicitly record the variables in the notes using standard definitions, perhaps on a clerking form.[32]

This study is important because we have used large samples to establish the reliability of the variables in our models in the environments in which they might realistically be used. However, the study also has certain shortcomings. Firstly, we performed two of our three assessments in the population of only one hospital and secondly, our data collection was performed either by trainee neurologists with an interest in stroke or by an experienced audit assistant. Whether these high levels of reliability and validity can be replicated in other populations and by other, less experienced observers remains to be established. Thirdly, our study has not considered the inter-rater reliability of the variables included in our models in patients with hyper-acute stroke. However, few such patients were included in the cohorts used to develop and test our models and therefore the relevance of our models to hyper-acute stroke also remains uncertain. Fourthly, it might be argued that familiarity led us to achieve higher levels of agreement for the variables included in our own models than for those included in models developed by others. However, beyond those encountered in daily clinical practice, neither NW nor CC had collected our model variables before the study. Lastly, our study has not compared the predictive accuracy of our models with that of the other models studied. However, this was not the aim of our study and, to be informative, would have required a much larger sample size; furthermore, the predictive accuracy of all the models in this study has been studied previously.

In conclusion, this study suggests that the variables in our simple models of outcome after stroke are very reliable when prospectively collected and reasonably reliable and valid when retrospectively collected. It is probable that the reliability of data collection would be improved if our variables were more explicitly defined and, for retrospective purposes, explicitly recorded in the routine medical record.

ACKNOWLEDGEMENTS

We are grateful to the staff and patients who participated in the Scottish Stroke Outcomes Study Group and to all those who contributed to the stroke register at the Western General Hospital, Edinburgh.

Authors' affiliations
N U Weir, C E Counsell, M McDowall, A Gunkel, M S Dennis, Department of Clinical Neurosciences, Western General Hospital, Edinburgh, UK

Funding: The Wellcome Trust funded Dr Nicolas Weir and Dr Carl Counsell as well as the original development of the prognostic models. The Stroke Outcomes Study was funded by the Scottish Chief Scientist Office and the Clinical Resource & Audit Group (CRAG).

Competing interests: none declared.

REFERENCES

1 **Kwakkel G**, Wagenaar RC, Kollen BJ, *et al*. Predicting disability in stroke–a critical review of the literature. *Age Ageing* 1996;**25**:479–89.
2 **Counsell C**, Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovasc Dis* 2001;**12**:159–70.
3 **Laupacis A**, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;**277**:488–94.
4 **Wyatt JC**, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? *BMJ* 1995;**311**:1539–41.
5 **Lindley RI**, Warlow CP, Wardlaw JM, *et al*. Interobserver reliability of a clinical classification of acute cerebral infarction. *Stroke* 1993;**24**:1801–4
6 **D'Olhaberriague L**, Litvan I, Mitsias P, *et al*. A reappraisal of reliability and validity studies in stroke. *Stroke* 1996;**27**:2331–6.
7 **Dewey HM**, Donnan GA, Freeman EJ, *et al*. Interrater reliability of the National Institutes of Health Stroke Scale: rating by neurologists and nurses in a community-based stroke incidence study. *Cerebrovasc Dis* 1999;**9**:323–7.
8 **Gelmers HJ**, Gorter K, de Weerdt CJ, *et al*. Assessment of interobserver variability in a Dutch multicenter study on acute ischemic stroke. *Stroke* 1988;**19**:709–11.
9 **Frithz G**, Werner I. Studies on cerebrovascular strokes. II. Clinical findings and short term prognosis in a stroke material. *Acta Med Scand* 1976;**199**:133–40.
10 **Romm FJ**, Putnam SM. The validity of the medical record. *Med Care* 1981;**19**:310–15.
11 **Osbourne CE**, Thompson HC. Criteria for evaluation of ambulatory child health care by chart audit: development and testing of a methodology. Final report of the Joint Committee on Quality Assurance of Ambulatory Health Care for Children and Youth. *Pediatrics* 1975;**56** (part 4, suppl 2):625–92.
12 **Caplan RA**, Posner KL, Cheney FW. Effect of outcome on physician judgments of appropriateness of care. *JAMA* 1991;**265**:1957–60.
13 **Goldstein LB**, Chilukuri V. Retrospective assessment of initial stroke severity with the Canadian Neurological Scale. *Stroke* 1997;**28**:1181–4.

14 **Baird AE**, Dashe J, Connor A, *et al*. Comparison of retrospective and prospective measurements of the National Institutes of Health Stroke Scale. *Cerebrovasc Dis* 2000;**10**:80–1.
15 **Kasner SE**, Chalela JA, Luciano JM, *et al*. Reliability and validity of estimating the NIH Stroke Scale from medical records. *Stroke* 1999;**30**:1534–7.
16 **Williams LS**, Yilmaz EY, Lopez-Yunez AM. Retrospective assessment of initial stroke severity with the NIH Stroke Scale. *Stroke* 2000;**31**:858–62.
17 **Bushnell CD**, Johnston DCC, Goldstein LB. Retrospective assessment of initial stroke severity. *Stroke* 2001;**32**:656–60.
18 **Counsell C**, Dennis M, McDowall M, *et al*. Predicting outcome after acute stroke: development and validation of new prognostic models. *Stroke* 2002;**33**:1041–7.
19 **Weir N**, Dennis MS for the Scottish Stroke Outcomes Study Group. Towards a national system for monitoring the quality of hospital based stroke services. *Stroke* 2001;**32**:1415–21.
20 **The FOOD Trial Collaboration**. Performance of a statistical model to predict stroke outcome in the context of a large simple randomised controlled trial of feeding. *Stroke* 2003;**34**:127–33.
21 **Hand P**, Lindley R, Wardlaw J, *et al*. The Third International Stroke Trial (IST3). [Abstract]. *Cerebrovasc Dis* 2001:**11**:35.
22 **Dennis M**. CLOTS (Clots in Legs Or TEDS after stroke). A randomised trial to establish the effectiveness of graduated compression stockings to prevent post stroke deep vein thrombosis (DVT). [Abstract]. *Cerebrovasc Dis* 2001:**11**:32.
23 **Gladman JR**, Harwood DM, Barer DH. Predicting the outcome of acute stroke: prospective evaluation of five multivariate models and comparison with simple methods. *J Neurol Neurosurg Psychiatry* 1992;**55**:347–51.
24 **Wade DT**, Hewer RL. Outlook after an acute stroke: urinary incontinence and loss of consciousness compared in 532 patients. *Q J Med* 1985;**56**:601–8.
25 **Allen CM**. Predicting the outcome of acute stroke: a prognostic score. *J Neurol Neurosurg Psychiatry* 1984;**47**:475–80.
26 **Prescott RJ**, Garraway WM, Akhtar AJ. Predicting functional outcome following acute stroke using a standard clinical examination. *Stroke* 1982;**13**:641–7.
27 **Kalra L**, Crome P. The role of prognostic scores in targeting stroke rehabilitation in elderly patients. *J Am Geriatr Soc* 1993;**41**:396–400.
28 **Altman DG**. Some common problems in medical research. In: *Practical statistics for medical research*. 1st edn.. London: Chapman and Hall, 1993:396–439.
29 **Fleiss JL**, Cohen J, Everitt BS. Large sample standard errors of Kappa and weighted Kappa. *Psychol Bull* 1969;**72**:323–7.
30 **Brennan P**, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;**304**:1491–4.
31 **Wade DT**. Personal physical disability. In: *Measurement in neurological rehabilitation*. Oxford: Oxford University Press, 1995:70–82.
32 **Davenport RJ**, Dennis MS, Warlow CP. Improving the recording of clinical assessment of stroke patients using a clerking proforma. *Age Ageing* 1995;**24**:43–8.